Nonconvergence to Saddle Boundary Points under Perturbed Reinforcement Learning[*]

Georgios C. Chasparis[†] Jeff S. Shamma[‡] Anders Rantzer[§]

December 7, 2012

Abstract

This paper presents a novel reinforcement learning algorithm and provides conditions for global convergence to Nash equilibria. For several classes of reinforcement learning schemes, including the ones proposed here, excluding convergence to action profiles which are not Nash equilibria may not be trivial, unless the step-size sequence is appropriately tailored to the specifics of the game. In this paper, we sidestep these issues by introducing a perturbed reinforcement learning scheme where the strategy of each agent is perturbed by a strategy-dependent perturbation (or mutations) function. Contrary to prior work on equilibrium selection in games where perturbation functions are globally state dependent, the perturbation function here is assumed to be local, i.e., it only depends on the strategy of each agent. We provide conditions under which the strategies of the agents will converge to an arbitrarily small neighborhood of the set of Nash equilibria almost surely. This extends prior analysis on reinforcement learning in games which has been primarily focused on urn processes. We finally specialize the results to a class of potential games.

1 Introduction

Lately, agent-based modeling has generated significant interest in various settings, such as engineering, social sciences and economics. In those formulations, agents make decisions independently and without knowledge of the actions or intentions of the other agents. Usually, the interactions among agents can be described in terms of a strategic-form game, and stability notions, such as the Nash equilibrium, can be utilized to describe desirable outcomes for all agents.

In this paper, we are interested in deriving conditions under which agents *learn* to play Nash equilibria. Assuming minimum information available to each agent, namely its own utility and actions, we introduce a

[*] This work was supported by the Swedish Research Council through the Linnaeus Center LCCC and the AFOSR MURI project #FA9550-09-1-0538.

[†] G.C. Chasparis is with the Software Competence Center, Hagenberg, Austria, gchasparis@gmail.com, http://www.chasparis.blogspot.gr

[‡] J.S. Shamma is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, shamma@gatech.edu, www.prism.gatech.edu/˜jshamma3

[§] A. Rantzer is with the Department of Automatic Control, Lund University, Lund, Sweden, rantzer@control.lth.se, www.control.lth.se/Staff/anders_ rantzer.html

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **07 DEC 2012**	2. REPORT TYPE	3. DATES COVERED **00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE
Nonconvergence to Saddle Boundary Points under Perturbed Reinforcement Learning

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Georgia Institute of Technology,School of Electrical and Computer Engineering,Atlanta,GA,30332

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR’S ACRONYM(S)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
submitted for journal publication

14. ABSTRACT

This paper presents a novel reinforcement learning algorithm and provides conditions for global convergence to Nash equilibria. For several classes of reinforcement learning schemes, including the ones proposed here, excluding convergence to action profiles which are not Nash equilibria may not be trivial, unless the step-size sequence is appropriately tailored to the specifics of the game. In this paper we sidestep these issues by introducing a perturbed reinforcement learning scheme where the strategy of each agent is perturbed by a strategy-dependent perturbation (or mutations) function. Contrary to prior work on equilibrium selection in games where perturbation functions are globally state dependent, the perturbation function here is assumed to be local, i.e., it only depends on the strategy of each agent. We provide conditions under which the strategies of the agents will converge to an arbitrarily small neighborhood of the set of Nash equilibria almost surely. This extends prior analysis on reinforcement learning in games which has been primarily focused on urn processes. We finally specialize the results to a class of potential games.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**30**	

novel reinforcement learning scheme and derive conditions under which global convergence to Nash equilibria can be achieved.

In reinforcement learning schemes, agents build their confidence over an action through repeated selection of this action and proportionally to the reward received from this action. In particular, each agent updates a probability distribution over its available actions and the probability of selecting an action increases whenever this action is selected and proportionally to the reward received. This class of dynamics has been applied in evolutionary economics, for modeling human and economic behavior [1, 2, 3, 4, 5] and sociology, for modeling social network formation [6, 7].

Reinforcement learning schemes are also related to replicator dynamics [8] as has been pointed out by several authors [2, 4, 5]. For example, in [2, 9], the asymmetric, continuous replicator dynamics (cf., [10]) have been identified as the continuous time limit of a reinforcement learning scheme which is based on Bush-Mosteller's [11] simple learning model.

One of the main concerns in the analysis of reinforcement learning schemes is showing nonconvergence to boundary points of the probability simplex which do not correspond to Nash equilibria. In fact, as pointed out in [4], establishing nonconvergence to the boundary of the probability simplex might not be trivial, since standard results of the ODE method for stochastic approximations (e.g., nonconvergence to unstable equilibria [12]) are not applicable. Thus, the behavior of several reinforcement learning models, e.g., the model by [1], cannot be directly related to (standard) replicator dynamics. This is mainly due to the fact that several models of reinforcement learning may converge to saddle boundary points of the replicator dynamics [4].

In this paper, we sidestep these issues by introducing a new class of reinforcement learning schemes where the strategies of each agent are perturbed by a state-dependent perturbation function. Contrary to prior work on equilibrium selection where perturbation functions are also state dependent [13], the perturbation function here is assumed to be local, i.e., it only depends on the strategy of each player. Due to this perturbation function, the ODE method for stochastic approximations can be applied, since boundary points of the domain cease to be stationary points of the relevant ODE. This paper extends prior work [14] of the authors, where the perturbation function was assumed constant independently of the strategy. In particular, we provide conditions under which the strategies of the agents will converge to an arbitrarily small neighborhood of the set of Nash equilibria almost surely.

We further specialize the results to a class of games which belongs to the family of *potential games* [15]. It includes *common-payoff* (or *identical-interest*) games, *congestion* games [16], and *two-player rescaled partnership* games [10]. Potential games are also of particular interest in engineering, for example in congestion control [16], distributed spatial coverage [17] and distributed routing [18]. In these examples, and when agents are playing the game repeatedly, learning to play a Nash equilibrium is of special interest, especially when the information available to each agent is only the history of its own utilities and its own actions. We provide conditions under which the proposed reinforcement learning scheme converges to the set of pure Nash equilibria for this class of games. This is also an extension of prior work on reinforcement learning [5, 4] in potential games which has primarily focused on the urn process of [3].

The remainder of the paper is organized as follows. Section 2 introduces the necessary terminology.

Section 3 introduces the perturbed reinforcement learning scheme with a state-based perturbation function. Section 4 states some standard results from Lyapunov-based techniques and the ODE method for analyzing stochastic approximations. Sections 5 characterizes the set of stationary points of the reinforcement learning scheme for both the unperturbed and the perturbed dynamics. Section 6 analyzes the behavior of the unperturbed reinforcement learning scheme close to the boundary points of the domain, while Section 7 analyzes the convergence properties of the perturbed learning scheme. Finally, Section 8 specializes the results to a class of games which belongs to the family of potential games, and Section 9 presents concluding remarks.

Notation:

– $|x|$ denotes the Euclidean norm of a vector $x \in \mathbb{R}^n$.

– $|x|_\infty$ denotes the ℓ_∞-norm of a vector $x \in \mathbb{R}^n$.

– $\mathcal{B}_\delta(x)$ denotes the δ-neighborhood of vector $x \in \mathbb{R}^n$, i.e.,

$$\mathcal{B}_\delta(x) \triangleq \{y \in \mathbb{R}^n : |x-y| < \delta\}.$$

– $\mathrm{dist}(x, A)$ from a point x to a set A is defined as

$$\mathrm{dist}(x, A) \triangleq \inf_{y \in A} |x-y| \, .$$

– $\mathcal{B}_\delta(A)$ denotes the δ-neighborhood of set $A \subset \mathbb{R}^n$, i.e.,

$$\mathcal{B}_\delta(A) \triangleq \{x : \mathrm{dist}(x, A) < \delta\}.$$

– $\Delta(m)$ denotes the probability simplex of dimension m, i.e.,

$$\Delta(m) \triangleq \left\{x \in \mathbb{R}^m : x \geq 0, \mathbf{1}^{\mathrm{T}}x = 1\right\},$$

where $\mathbf{1}$ is the vector of ones of appropriate dimension.

– $\Pi_\Delta : \mathbb{R}^m \to \Delta(m)$ is the projection to the probability simplex, i.e.,

$$\Pi_\Delta[x] \triangleq \arg \min_{y \in \Delta(m)} |x-y|.$$

– A^o is the interior of a subset A of $\mathbb{R}^n$, and ∂A is its boundary.

– $\mathrm{row}\{\alpha_i\}_{i\in\mathcal{J}}$ denotes the block row vector with entries $\{\alpha_i\}_{i\in\mathcal{J}}$ for some set of indices $\mathcal{J}$, i.e.,

$$\mathrm{row}\{\alpha_i\}_{i\in\mathcal{J}} \triangleq \begin{pmatrix} \alpha_1 & \cdots & \alpha_{|\mathcal{J}|} \end{pmatrix},$$

where $\alpha_i \in \mathbb{R}^{1\times n_i}$ for some $n_i \in \mathbb{N}$, $i \in \mathcal{J}$. Likewise, $\mathrm{col}\{\cdot\}$ will denote a block column vector.

– $\mathrm{diag}\{A_i\}_{i\in\mathcal{J}}$ denotes the block diagonal matrix with diagonal entries $\{A_i\}_{i\in\mathcal{J}}$ for some set of indices $\mathcal{J}$, i.e.,

$$\mathrm{diag}\{A_i\}_{i\in\mathcal{J}} \triangleq \begin{pmatrix} A_1 & & \\ & \ddots & \\ & & A_{|\mathcal{J}|} \end{pmatrix},$$

where $A_i \in \mathbb{R}^{n_i \times m_i}$ for some $n_i, m_i \in \mathbb{N}$, $i \in \mathcal{J}$.

2 Terminology

We consider the standard setup of finite strategic-form games.

2.1 Game

A finite strategic-form game involves a finite set of *agents* (or *players*), $\mathcal{I} \triangleq \{1, 2, ..., n\}$. Each agent $i \in \mathcal{I}$ has a *finite* set of available *choices* (or *actions*), $\mathcal{A}_i$. Let $\alpha_i \in \mathcal{A}_i$ denote an *action* of agent i, and $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)$ the *action profile* of all agents. The set $\mathcal{A}$ is the Cartesian product of the action spaces of all agents, i.e., $\mathcal{A} \triangleq \mathcal{A}_1 \times ... \times \mathcal{A}_n$.

The action profile $\alpha \in \mathcal{A}$ produces a *payoff* (or *utility*) for each agent. The utility of agent i, denoted by R_i, is a function which maps the action profile α to a payoff in $\mathbb{R}$. It constitutes a measure of the desirability of the action profile α, where a high-payoff action profile is more desirable than a low-payoff action profile. Let also denote by $R : \mathcal{A} \to \mathbb{R}^n$ the combination of payoffs (or *payoff profile*) of all agents, i.e., $R(\cdot) \triangleq (R_1(\cdot), R_2(\cdot), ..., R_n(\cdot))$. A *strategic-form game* will then be completely characterized by the triple $\{\mathcal{I}, \mathcal{A}, R\}$.

2.2 Strategy

Since each agent selects actions independently, we generally assume that each agent's action is a realization of an independent discrete random variable. Let $\sigma_{ij} \in [0, 1]$ denote the probability that agent i selects action $\alpha_i = j \in \mathcal{A}_i$. If $\sum_{j\in\mathcal{A}_i} \sigma_{ij} = 1$, then $\sigma_i \triangleq (\sigma_{i1}, \sigma_{i2}, ..., \sigma_{i|\mathcal{A}_i|})$ is a probability distribution over the set of actions $\mathcal{A}_i$ (or *strategy* of agent i), where $|\mathcal{A}_i|$ denote the cardinality of the set $\mathcal{A}_i$. Then $\sigma_i \in \Delta(|\mathcal{A}_i|)$. We will also use the term *strategy profile* to denote the combination of strategies of all agents $\sigma = (\sigma_1, \sigma_2, ..., \sigma_n) \in \boldsymbol{\Delta}$ where $\boldsymbol{\Delta} \triangleq \Delta(|\mathcal{A}_1|) \times ... \times \Delta(|\mathcal{A}_n|)$ is the set of strategy profiles.

Note that if σ_i is a *unit vector* (or a vertex of $\Delta(|\mathcal{A}_i|)$), say e_j, then agent i selects an action j with probability one. Such a strategy will be called *pure strategy*. Likewise, a *pure strategy profile* is a profile of pure strategies. We will also use the term *mixed strategy* to denote a strategy that is *not* pure.

2.3 Expected payoff and Nash equilibrium

Given a strategy profile $\sigma \in \boldsymbol{\Delta}$, the *expected payoff vector* of each agent i, $U_i : \boldsymbol{\Delta} \to \mathbb{R}^{|\mathcal{A}_i|}$, can be computed by

$$U_i(\sigma) \triangleq \sum_{j\in\mathcal{A}_i} e_j \sum_{\alpha_{-i}\in\mathcal{A}_{-i}} \left(\prod_{s\in -i} \sigma_{s\alpha_s}\right) R_i(j, \alpha_{-i}). \tag{1}$$

We may think of the entry j of the expected payoff vector, denoted $U_{ij}(\sigma)$, as the payoff of agent i who is playing action j at strategy profile σ. We denote the profile of expected payoffs by $U(\sigma) = (U_1(\sigma), ..., U_n(\sigma))$. Finally, let $u_i(\sigma)$ be the *expected payoff* of agent i at strategy profile $\sigma \in \boldsymbol{\Delta}$, defined as follows:

$$u_i(\sigma) \triangleq \sigma_i^{\mathrm{T}} U_i(\sigma).$$

In the trivial case of $n = 2$, it is straightforward to check that for every $i \in \mathcal{I}$, there exists matrix $D_i \in \mathbb{R}^{|\mathcal{A}_i|\times|\mathcal{A}_{-i}|}$, such that $D_i = [R_i(j,\ell)]_{j\ell}$.[1] In this case, the expected payoff of player i can be written in the simplified form:

$$u_i(\sigma) = \sigma_i^{\mathrm{T}} D_i \sigma_{-i}.$$

Definition 2.1 (Nash equilibrium) *A strategy profile $\sigma^* = (\sigma_1^*, \sigma_2^*, ..., \sigma_n^*) \in \boldsymbol{\Delta}$ is a Nash equilibrium if, for each agent $i \in \mathcal{I}$,*

$$u_i(\sigma_i^*, \sigma_{-i}^*) \geq u_i(\sigma_i, \sigma_{-i}^*) \tag{2}$$

for all $\sigma_i \in \Delta(|\mathcal{A}_i|)$ and $\sigma_i \neq \sigma_i^*$, where σ_{-i}^* denote the equilibrium strategy profile of all agents but i.

In the special case where for all $i \in \mathcal{I}$, σ_i^* is a pure strategy, then the Nash equilibrium is called *pure Nash equilibrium*. Also, in case the inequality in (2) is strict, the Nash equilibrium will be called a *strict Nash equilibrium*.

3 Perturbed Learning Automata

In this section, we introduce the basic form of reinforcement learning that we will consider in the remainder of the paper. It belongs to the general class of *learning automata* [19].

The basic idea behind a reinforcement learning scheme is a rather simple one. If agent i selects action j at instant k and a favorable payoff results, $R_i(\alpha(k))$, the action probability $\sigma_{ij}(k)$ is increased and all the other components of $\sigma_i(k)$ are decreased. For an unfavorable payoff, $\sigma_{ij}(k)$ is decreased and all the other components of $\sigma_i(k)$ are increased.

The precise manner in which $\sigma_i(k)$ is changed depending on the action α_i performed at stage k and the response $R_i(\alpha(k))$ of the environment, completely defines the reinforcement scheme. This, in turn, determines the resulting Markov process and hence the behavior of the overall system.

For the remainder of the paper, we will assume:

[1] The notation $-i$ denotes the complementary set $\mathcal{I}\backslash\{i\}$. We will often split the argument of a function in this way, e.g., $F(\alpha) = F(\alpha_i, \alpha_{-i})$ or $F(x) = F(x_i, x_{-i})$.

Assumption 3.1 (Strictly positive rewards) *For every $i \in \mathcal{I}$, the reward function satisfies $R_i(\alpha) > 0$ for all $\alpha \in \mathcal{A}$.*

3.1 Modified Linear Reward-Inaction ($\widetilde{\mathcal{L}}_{\mathrm{R-I}}$) scheme

We consider a reinforcement scheme which is a small modification of the original *linear reward-inaction scheme* ($\mathcal{L}_{\mathrm{R-I}}$) introduced by [20, 21]. This modified scheme, denoted by $\widetilde{\mathcal{L}}_{\mathrm{R-I}}$, was introduced in [14]. Compared with $\mathcal{L}_{\mathrm{R-I}}$, the reward in $\widetilde{\mathcal{L}}_{\mathrm{R-I}}$ may take values other than $\{0, 1\}$, which increases the family of games that this learning scheme can be applied to.

Similarly to $\mathcal{L}_{\mathrm{R-I}}$, the probability that agent i selects action j at time k is

$$\sigma_{ij}(k) = x_{ij}(k)$$

for some probability vector $x_i(k)$ which is updated according to the recursion:

$$x_i(k+1) = \Pi_{\Delta}\left[x_i(k) + \epsilon(k) \cdot R_i(\alpha(k)) \cdot [e_{\alpha_i(k)} - x_i(k)]\right]. \tag{3}$$

Here we identify actions $\mathcal{A}_i$ with vertices of the simplex, $\{e_1, ..., e_{|\mathcal{A}_i|}\}$. For example, if agent i selects action j at time k, then $e_{\alpha_i(k)} = e_j$. Note that by letting the step-size sequence $\epsilon(k)$ to be sufficiently small and since the payoff function $R_i(\cdot)$ is uniformly bounded in $\mathcal{A}$, $x_i(k) \in \Delta(|\mathcal{A}_i|)$ and the projection operator can be omitted.

We consider the following class of step-size sequences:

$$\epsilon(k) = \frac{1}{k^{\nu} + 1} \tag{4}$$

for some $\nu \in (1/2, 1]$. For these values of ν, the following two conditions can easily be verified:

$$\sum_{k=0}^{\infty} \epsilon(k) = \infty \quad \text{and} \quad \sum_{k=0}^{\infty} \epsilon(k)^2 < \infty. \tag{5}$$

The selection of ν is closely related to the desired rate of convergence.

Compared with prior reinforcement schemes, in particular the models of [1, 4], the main difference lies in the step-size sequence. More specifically, in [1] the step-size sequence of agent i is $\epsilon_i(k) = 1/(ck^{\nu} + R_i(\alpha))$ for some positive constant c and for $0 < \nu < 1$. A comparative model is also used by [4] with a step-size sequence to be $\epsilon_i(k) = 1/(V_i(k) + R_i(\alpha(k)))$, where $V_i(k)$ is the accumulated benefits of agent i up to time k, which gives rise to an urn process introduced by [3]. Some similarities are also shared with the Cross' learning model of [2], where $\epsilon(k) = 1$ and $R_i(\alpha(k)) \leq 1$, and its modification presented in [9], where $\epsilon(k)$, instead, is assumed decreasing. The aforementioned reinforcement schemes do not have identical convergence properties with $\widetilde{\mathcal{L}}_{\mathrm{R-I}}$. Their differences will be discussed in detail throughout the paper.

3.2 Pertubed Linear Reward-Inaction Scheme ($\widetilde{\mathcal{L}}^{\lambda}_{\mathrm{R-I}}$)

Here we consider a perturbed version of the $\widetilde{\mathcal{L}}_{\mathrm{R-I}}$ scheme, in the same spirit with [14], where the decisions of each agent are slightly perturbed. In particular, we assume that each agent i selects action $j \in \mathcal{A}_i$ according to the perturbed strategy

$$\sigma_{ij} \triangleq (1-\zeta_i(x_i,\lambda))x_{ij} + \zeta_i(x_i,\lambda)/\left|\mathcal{A}_i\right|, \tag{6}$$

for some perturbation function $\zeta_i : \Delta(|\mathcal{A}_i|) \times [0,1] \to [0,1]$ (usually called *mutations*).

The introduction of a state-dependent mutations function intends on exploring how local information of each agent i, namely its state x_i, can alter the convergence properties of the state of the group. Here, we investigate one class of such mutations function. In particular, we would like $\zeta_i(x_i,\lambda)$ to exhibit larger values when the strategy x_i is close to a vertex of the probability simplex, i.e., the strategy is close to a pure strategy. Informally, players are "exploring" the most when they are "certain" about which action to choose.

Formally, we will consider the following class of perturbation functions:

Assumption 3.2 (Perturbation function) *The perturbation function ζ_i is continuously differentiable. Furthermore, for some $\beta \in (0,1)$ sufficiently close to one, ζ_i satisfies the following properties:*

1. $\zeta_i(x_i,\lambda) = 0$ *for all x_i such that $|x_i|_\infty < \beta$ for any $\lambda \geq 0$;*
2. $\lim_{|x_i|_\infty \to 1} \zeta_i(x_i,\lambda) = \lambda$*;*
3. $\lim_{|x_i|_\infty \to 1} \frac{\partial \zeta_i(x_i,\lambda)}{\partial \lambda}|_{(\lambda=0)} = c$ *for some $c > 0$;*
4. $\lim_{|x_i|_\infty \to 1} \frac{\partial \zeta_i(x_i,\lambda)}{\partial x_{ij}}|_{(\lambda=0)} = 0$ *for any $j \in \mathcal{A}_i$.*

In other words, we would like the perturbation function of agent i (1) to be zero when its strategy is not close to a vertex of $\Delta(|\mathcal{A}_i|)$; and (2) to be equal to λ when its strategy is at a vertex of $\Delta(|\mathcal{A}_i|)$. Properties (3) and (4) are necessary in order to analyze the behavior of the stochastic process in the vicinity of the vertices of $\Delta(|\mathcal{A}_i|)$. In particular, property (3) states that the perturbation increases with λ, when evaluated at a vertex of the probability simplex and for $\lambda = 0$. As we shall see in a forthcoming section, due to this property, vertices cease to be stationary points of the mean-field dynamics introduced in Section 4, which has favorable implications on the asymptotic behavior of the learning dynamics. Finally, property (4) states that the perturbation does not change with x when evaluated at a vertex of the probability simplex and for $\lambda = 0$. Together with property (1), property (4) establishes equivalence among perturbed and unperturbed dynamics when $\lambda = 0$.

For example, a candidate perturbation function is:

$$\zeta_i(x_i,\lambda) = \begin{cases} 0 & |x_i|_\infty < \beta, \\ \frac{\lambda}{(1-\beta)^2}(|x_i|_\infty - \beta)^2 & |x_i|_\infty \geq \beta. \end{cases} \tag{7}$$

It is straightforward to check that this function satisfies the properties of Assumption 3.2 when we select $\beta \in (0,1)$ sufficiently close to one. Figure 1 plots the candidate perturbation function (7) about one of the

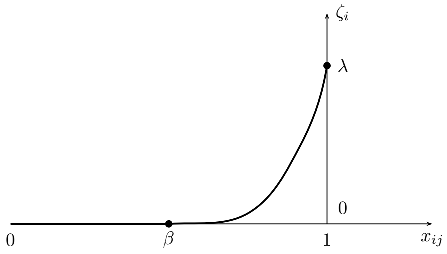


Figure 1: Candidate perturbation function (7).

vertices of the domain $\Delta(|\mathcal{A}_i|)$.

Note that the main difference with the previously introduced scheme in [14] is that here we allow for the perturbation function of each agent to also depend on agent's own strategy. Instead, in [14], the perturbation function was assumed to be a constant $\lambda > 0$ for all agents $i \in \mathcal{I}$ and for all strategy vectors in $\Delta(|\mathcal{A}_i|)$.

We will denote this scheme by $\widetilde{\mathcal{L}}^{\lambda}_{\mathrm{R-I}}$.

3.3 Discussion

Similar ideas of state dependent mutations have been explored in aspiration learning [22] and in adaptive learning state-dependent excitation of the dynamics is to establish an [13]. In both references, the intention of a equilibrium selection scheme that will give rise to more desirable outcomes. For example, in [22], the intention is to show that the aspiration learning scheme will converge to a Pareto efficient action profile. In a similar spirit, reference [13] introduces globally state dependent mutations to show that each action profile can be a stochastically stable outcome of an evolutionary learning process, when we tailor appropriately the mutations function.

Our intention here is to also use the state-dependent perturbation function as an equilibrium selection mechanism. In comparison with [22], our goal is to analyze the asymptotic behavior of a class of reinforcement learning schemes, whose behavior is quite different than aspiration learning. In comparison with [13], our class of perturbation functions for each agent i are also state dependent, however they only depend on the strategy of each agent i and not on the strategy profile of all agents.

Furthermore, the introduction of such perturbation function serves as an alternative scheme for analyzing convergence to boundary points of the probability simplex compared to prior analysis in both [1] and [4]. In particular, as [4] points out, the behavior of general models of reinforcement learning, such as the model by [1], cannot be directly related to standard replicator dynamics (cf., [10, Chapter 7]). This is mainly due to the fact that several models of reinforcement learning may converge to saddle points of the standard replicator dynamics. As it will become clear later on, such issues will be sidestepped here due to the introduction of the mutations function of $\widetilde{\mathcal{L}}^{\lambda}_{\mathrm{R-I}}$.

4 Background Convergence Analysis

Let $\Omega \triangleq \mathbf{\Delta}^\infty$ denote the canonical path space with an element ω being a sequence $\{x(0), x(1), ...\}$, where $x(k) = (x_1(k), ..., x_n(k)) \in \mathbf{\Delta}$ is generated by the reinforcement learning process. An example of a random variable defined in Ω is the function $\psi_k : \Omega \to \mathbf{\Delta}$ such that $\psi_k(\omega) = x(k)$. Another example of a random variable that we will also use is $\psi_k(\omega) = \alpha(k)$. In several cases, we will abuse notation by simply writing $x(k)$ or $\alpha(k)$ instead of $\psi_k(\omega)$. Let also $\mathcal{F}$ be a σ-algebra of subsets in Ω and $\mathbb{P}$, $\mathbb{E}$ be the probability and expectation operator on $(\Omega, \mathcal{F})$, respectively. In the following analysis, we implicitly assume that the σ-algebra $\mathcal{F}$ is generated appropriately to allow computation of the probabilities or expectations of interest.

To analyze the asymptotic behavior of the reinforcement learning schemes, we will use a) stochastic Lyapunov stability analysis, in order to investigate the probabilities that a sample function exits from a domain, and b) the ODE method for stochastic approximations in order to investigate the probability of convergence to invariant sets of the mean-field dynamics. The background analysis which is necessary for the analysis are presented in the following subsections.

4.1 Exit of a sample function from a domain

It is important to have conditions under which the process $\psi_k(\omega) = x(k)$, $k \geq 0$, with some initial distribution, will exit an open domain G in finite time.

Proposition 4.1 (Theorem 5.1 in [23]) *Suppose that there exists a nonnegative function, $V(k, x)$ in the domain $k \geq 0$, $x \in G$, such that*

$$\Delta V(k, x) \triangleq \mathbb{E}[V(k+1, x(k+1)) - V(k, x(k)) | x(k) = x]$$

satisfies $\Delta V(k, x) \leq -a(k)$ in this domain, where $a(k)$ is a sequence such that

$$a(k) > 0, \quad \sum_{k=0}^{\infty} a(k) = \infty. \tag{8}$$

Then, the process $x(k)$ leaves G in a finite time with probability 1.

The following corollary is important in cases we would like to consider entrance of a stochastic process into the domain of attraction of an equilibrium. It is a direct consequence of Proposition 4.1. For details, see Exercise 5.1 in [23].

Corollary 4.1 *Let $A \subset \mathbf{\Delta}$, $\mathcal{B}_\delta(A)$ its δ-neighborhood, and $\mathcal{D}_\delta(A) \triangleq \mathbf{\Delta} \backslash \mathcal{B}_\delta(A)$. Suppose there exists a nonnegative function $V(k, x)$ in the domain $k \geq 0$, $x \in \mathbf{\Delta}$ for which*

$$\Delta V(k, x) \leq -a(k)\varphi(k, x), \quad k \geq 0, x \in \mathbf{\Delta}, \tag{9}$$

where the sequence $a(k)$ satisfies (8) and $\varphi(k,x)$ satisfies

$$\inf_{k\geq T, x\in\mathcal{D}_\delta(A)} \varphi(k,x) > 0$$

for all $\delta > 0$ and some $T = T(\delta)$. Then

$$\mathbb{P}\Big[\liminf_{k\to\infty} \operatorname{dist}(x(k), A) = 0\Big] = 1.$$

Corollary 4.1 implies that the stochastic process enters an arbitrarily small neighborhood of a set A infinitely often with probability one.

4.2 Convergence to mean-field dynamics

The convergence properties of $\widetilde{\mathcal{L}}^{\lambda}_{\mathrm{R-I}}$ can be described via the ODE method for stochastic approximations. In particular, the recursion of $\widetilde{\mathcal{L}}^{\lambda}_{\mathrm{R-I}}$, $\lambda \geq 0$, can be written in the following form:

$$x_i(k+1) = x_i(k) + \epsilon(k)\cdot[\overline{g}^{\lambda}_i(x(k)) + \xi^{\lambda}_i(k)], \tag{10}$$

where the observation sequence has been decomposed into a deterministic sequence, $\overline{g}^{\lambda}_i(x(k))$, (or *mean-field*) and a noise sequence $\xi^{\lambda}_i(k)$. The mean-field is defined as follows:

$$\overline{g}^{\lambda}_i(x) \triangleq \mathbb{E}\left[R_i(\alpha(k))[e_{\alpha_i(k)} - x_i(k)] | x(k) = x\right]$$

such that its s-th entry is

$$\overline{g}^{\lambda}_{is}(x) = U_{is}(x)\sigma_{is} - \sum_{q\in\mathcal{A}_i} U_{iq}(x)\sigma_{iq}x_{is}.$$

It is straightforward to verify that $\overline{g}^{\lambda}_i(\cdot)$ is continuously differentiable due to the definition of the perturbation function ζ_i. The noise sequence is defined as

$$\xi^{\lambda}_i(k) \triangleq R_i(\alpha(k))\cdot\left[e_{\alpha_i(k)} - x_i(k)\right] - \overline{g}^{\lambda}_i(x(k)),$$

where $\mathbb{E}[\xi^{\lambda}_i(k)|x(k) = x] = 0$ for all $x \in \boldsymbol{\Delta}$.

Note that for $\lambda = 0$, (10) coincides with $\widetilde{\mathcal{L}}_{\mathrm{R-I}}$. We will denote $\overline{g}(x)$ the corresponding vector field.

The following more compact form of (10) also will be used:

$$x(k+1) = x(k) + \epsilon(k)\cdot\Big[\overline{g}^{\lambda}(x(k)) + \xi^{\lambda}(k)\Big], \tag{11}$$

where $\overline{g}^{\lambda}(\cdot) \triangleq \operatorname{col}\{\overline{g}^{\lambda}_i(\cdot)\}_{i\in\mathcal{I}}$ and $\xi^{\lambda}(\cdot) \triangleq \operatorname{col}\{\xi^{\lambda}_i(\cdot)\}_{i\in\mathcal{I}}$.

Proposition 4.2 (Convergence) *For the reinforcement scheme $\widetilde{\mathcal{L}}^{\lambda}_{\mathrm{R-I}}$, $\lambda \geq 0$, the stochastic iteration (11) is such that, for almost all $\omega \in \Omega$, $\{\psi_k(\omega) = x(k)\}$ converges to some bounded invariant set of the ODE:*

$$\dot{x} = \overline{g}^{\lambda}(x). \tag{12}$$

Furthermore, if $A \subset \mathbf{\Delta}$ is a locally asymptotically stable set in the sense of Lyapunov for (12),[2] *and $x(k)$ is in some compact set in the domain of attraction of A infinitely often with probability $\geq \rho$, then $x(k) \to A$ with at least probability ρ.*

Proof. The proposition follows from Theorem 6.6.1 of [24] since the following conditions are satisfied:

- The function $\overline{g}^{\lambda}(\cdot)$ is continuous.
- The sequence $Y(k) \triangleq \overline{g}^{\lambda}(x(k)) + \xi^{\lambda}(k)$ satisfies $\sup_k \mathbb{E}[|Y(k)|^2] < \infty$ since, by Assumption 3.1, the reward function is positive and bounded from above.
- The step size sequence satisfies $\sum_k \epsilon(k)^2 < \infty$ and $\sum_k \epsilon(k) = \infty$.

□ □

5 Stationary Points

Stationary points of the mean-field dynamics are defined as the set of points $x \in \mathbf{\Delta}$ for which $\overline{g}^{\lambda}(x) = 0$. In the following subsections, we characterize the set of stationary points for both the *unperturbed* ($\lambda = 0$) and the *perturbed* dynamics ($\lambda > 0$).

We will make the following distinction among stationary points of (12) for $\lambda > 0$, denoted $\mathcal{S}^{\lambda}$:

- $\mathcal{S}^{\lambda}_{\partial\mathbf{\Delta}}$: stationary points in $\partial\mathbf{\Delta}$;
- $\mathcal{S}^{\lambda}_{\mathbf{\Delta}^*}$: stationary points which are vertices of $\mathbf{\Delta}$;
- $\mathcal{S}^{\lambda}_{\mathbf{\Delta}^o}$: stationary points in $\mathbf{\Delta}^o$;
- $\mathcal{S}^{\lambda}_{\mathrm{NE}}$: stationary points which are Nash equilibria.

We will also use the notation $\mathcal{S}_{\partial\mathbf{\Delta}}$, $\mathcal{S}_{\mathbf{\Delta}^*}$, $\mathcal{S}_{\mathbf{\Delta}^o}$, and $\mathcal{S}_{\mathrm{NE}}$ to denote the corresponding sets when $\lambda = 0$.

5.1 Stationary Points of Unperturbed Dynamics ($\lambda = 0$)

Before describing the stationary points of the mean-field dynamics (12) under the unperturbed reinforcement learning ($\lambda = 0$), it is important to point out that the corresponding mean-field of the *share* of strategy s in agent i when $\lambda = 0$ can be written as:

$$\overline{g}_{is}(x) = \left(U_{is}(x) - \sum_{q \in \mathcal{A}_i} U_{iq}(x) x_{iq} \right) x_{is} \tag{13}$$

[2] If $\{x(t) : t \geq 0\}$ denotes the solution of the ODE (12), then a set $A \subset \mathbf{\Delta}$ is a *locally asymptotically stable set in the sense of Lyapunov* for the ODE (12) if a) for each $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that $\mathrm{dist}(x(0), A) < \delta$ implies $\mathrm{dist}(x(t), A) < \varepsilon$ for all $t \geq 0$, and b) there exists $\delta > 0$ such that $\mathrm{dist}(x(0), A) < \delta$ implies $\lim_{t\to\infty} \mathrm{dist}(x(t), A) = 0$.

which coincides with the corresponding shares provided by the standard replicator dynamics, as pointed out by [2, Proposition 1]. This form of dynamics can also be thought of as a special class of imitative dynamics, as discussed in [25, Section 5.4].

The following more compact form of standard replicator dynamics will be more convenient:

$$\overline{g}_i(x) = X_i(x_i) \cdot U_i(x), \quad i \in \mathcal{I}, \tag{14}$$

where $X_i : \Delta(|\mathcal{A}_i|) \to \mathbb{R}^{|\mathcal{A}_i| \times |\mathcal{A}_i|}$, such that $[X_i(x_i)]_{jj} = x_{ij}(1 - x_{ij})$ for any $j \in \mathcal{A}_i$ and $[X_i(x_i)]_{jk} = -x_{ij}x_{ik}$ for any $j, k \in \mathcal{A}_i$, with $j \neq k$.

The following proposition and corollaries characterize the stationary points of the ODE (12) for $\lambda = 0$ and are well known results for replicator dynamics (see, e.g., Section 3.3.1 in [26]).

Proposition 5.1 (Stationary points of unperturbed dynamics) *For $\lambda = 0$, a strategy profile x^* is a stationary point of the ODE (12) if and only if, for every agent $i \in \mathcal{I}$, there exists a constant $c_i > 0$, such that for any action $j \in \mathcal{A}_i$, $x^*_{ij} > 0$ implies $U_{ij}(x^*) = c_i$.*

Two straightforward implications of Proposition 5.1 are:

Corollary 5.1 (Pure Strategies) *For $\lambda = 0$, any pure strategy profile is a stationary point of the ODE (12).*

Corollary 5.2 (Nash Equilibria) *For $\lambda = 0$, any Nash equilibrium is a stationary point of the ODE (12).*

Note that for some games not all stationary points of the ODE (12) are Nash equilibria. For example, if you consider the Typewriter Game of Table 1, the pure strategy profiles which correspond to (A, B) or (B, A) are not Nash equilibria, although they are stationary points of (12).

	A	B
A	$4,4$	$2,2$
B	$2,2$	$3,3$

Table 1: The Typewriter Game.

On the other hand, any stationary point in the interior of the probability simplex will necessarily be a Nash equilibrium as the following corollary states:

Corollary 5.3 (Mixed Nash equilibria) *For $\lambda = 0$, any stationary point x^* of the ODE (12) for $\lambda = 0$, such that $x^* \in \mathbf{\Delta}^o$, is a (mixed) Nash equilibrium of the game.*

Note that the above corollaries do not exclude the possibility that there exist stationary points in $\partial\mathbf{\Delta}$ without those necessarily being pure strategy profiles. ***For the remainder of the paper***, we will only consider games which satisfy the following property:

Assumption 5.1 *For the unperturbed dynamics, there are no stationary points in $\partial\mathbf{\Delta}$ other than the ones in $\mathbf{\Delta}^*$, i.e., $\mathcal{S}_{\partial\mathbf{\Delta}} \backslash \mathcal{S}_{\mathbf{\Delta}^*} = \varnothing$. Moreover, if $\mathcal{S}_{\mathbf{\Delta}^o} \neq \varnothing$, there exists $\delta > 0$ such that $\mathcal{B}_\delta(\mathcal{S}_{\mathbf{\Delta}^o}) \subset \mathbf{\Delta}^o$.*

In other words, we only consider games for which, the stationary points of (12) for $\lambda = 0$ in the boundary of $\mathbf{\Delta}$ are vertices of $\mathbf{\Delta}$, and the stationary points in $\mathbf{\Delta}^o$ are isolated from the boundary. Assumption 5.1 is not restrictive and is satisfied for most but trivial cases.

Note also that Assumption 5.1 does not exclude the possibility that the vector field $\overline{g}(x)$ exhibits invariant sets other than stationary points.

5.2 Stationary Points of Perturbed Dynamics ($\lambda > 0$)

A straightforward implication of the properties of the perturbation function is the following:

Lemma 5.1 (Sensitivity of $S_{\mathbf{\Delta}^o}$) *There exists $\beta_0 \in (0, 1)$ such that $\mathcal{S}_{\mathbf{\Delta}^o} \subseteq \mathcal{S}^{\lambda}_{\mathbf{\Delta}^o}$ for any $\beta_0 < \beta < 1$ and any $\lambda > 0$.*

Proof. Due to Assumption 5.1, there exist $\beta_0 \in (0, 1)$ sufficiently close to one and $\delta > 0$, such that, for any $\beta_0 < \beta < 1$, we have $\zeta_i(x_i, \lambda) = 0$ for all $i \in \mathcal{I}$ and $x \in \mathcal{B}_\delta(\mathcal{S}_{\mathbf{\Delta}^o})$. Thus, the conclusion follows. □ □

Vertices of $\mathbf{\Delta}$ cease to be equilibria for $\lambda > 0$. The following proposition provides the sensitivity of $\mathcal{S}_{\mathbf{\Delta}^*}$ to small values of λ.

Lemma 5.2 (Sensitivity of $S_{\mathbf{\Delta}^*}$) *For any stationary point $x^* \in \mathcal{S}_{\mathbf{\Delta}^*}$, which corresponds to a strict Nash equilibrium and for sufficiently small $\lambda > 0$, there exists a unique continuously differentiable function $\nu^* : \mathbb{R}_+ \to \mathbb{R}^{|\mathcal{A}|}$, such that $\lim_{\lambda \downarrow 0} \nu^*(\lambda) = \nu^*(0) = 0$, and*

$$\tilde{x} = x^* + \nu^*(\lambda) \in \mathbf{\Delta}^o \tag{15}$$

is a stationary point of the ODE (12). If instead $x^* \in \mathcal{S}_{\mathbf{\Delta}^*}$ is not a Nash equilibrium, then for any sufficiently small $\delta > 0$ and $\lambda > 0$, the δ-neighborhood of x^* in $\mathbf{\Delta}$, $\mathcal{B}_\delta(x^*)$, does not contain any stationary point of the ODE (12).

Proof. The proof follows similar reasoning with the Proof of Proposition 3.5 in [14]. In the Appendix A, we present the main steps of the proof. □ □

Note that the statements of Lemma 5.2 do not depend on the selection of β. Instead, they require λ to be sufficiently small. Also, note that Lemma 5.2 does not discuss the sensitivity of Nash equilibria which are *not* strict. However, it is straightforward to show that vertices *cannot* be stationary points for $\lambda > 0$.

Let also $\widetilde{\mathcal{S}}^{\lambda}_{\mathrm{NE}}$ denote the set of stationary points in $\mathbf{\Delta}^o$ which are perturbations of the stationary points in $\mathcal{S}_{\mathbf{\Delta}^*} \cap \mathcal{S}_{\mathrm{NE}}$ (*strict* or *non-strict*) for some $\lambda > 0$.

Proposition 5.2 (Stationary points of perturbed dynamics) *For any $\beta \in (0, 1)$, let $\delta^* = \delta^*(\beta)$ be the smallest $\delta > 0$ such that, for all $x \in \mathbf{\Delta} \backslash \mathcal{B}_\delta(\mathbf{\Delta}^*)$, $\zeta_i(x_i, \lambda) = 0$ for some $i \in \mathcal{I}$. When β is sufficiently close to one and $\lambda > 0$ is sufficiently small, then:*

(a) $\widetilde{\mathcal{S}}^{\lambda}_{\mathrm{NE}} \subset \mathcal{B}_{\delta^*}(\mathbf{\Delta}^*)$;

(b) $\mathcal{S}^{\lambda} = \mathcal{S}_{\mathbf{\Delta}^o} \cup \widetilde{\mathcal{S}}^{\lambda}_{\text{NE}}$.

In other words, the stationary points of the perturbed dynamics are either the interior stationary points of the unperturbed dynamics or perturbations of pure Nash equilibria. Proposition 5.2 is an immediate implication of Lemmas 5.1–5.2 and Assumption 5.1. **Proof.** Pick $\beta > \beta_0$, where β_0 is defined in Lemma 5.1. Then $\mathcal{S}_{\mathbf{\Delta}^o} \subseteq \mathcal{S}^{\lambda}_{\mathbf{\Delta}^o} \equiv \mathcal{S}^{\lambda}$. The rest of the stationary points are perturbations of the vertices characterized by Lemma 5.2. Due to the definition of $\delta^* = \delta^*(\beta)$, we have $\widetilde{\mathcal{S}}^{\lambda}_{\text{NE}} \subset \mathcal{B}_{\delta^*}(\mathbf{\Delta}^*)$, since outside $\mathcal{B}_{\delta^*}(\mathbf{\Delta}^*)$ the dynamics coincide with the unperturbed dynamics for at least one agent. When we further take β to be sufficiently close to one (which implies that $\delta^* = \delta^*(\beta)$ approaches zero) and λ sufficiently small, then, according to Lemma 5.2, $\widetilde{\mathcal{S}}^{\lambda}_{\text{NE}}$ are the only stationary points in $\mathcal{B}_{\delta^*}(\mathbf{\Delta}^*)$, and therefore $\mathcal{S}^{\lambda} = \mathcal{S}_{\mathbf{\Delta}^o} \cup \widetilde{\mathcal{S}}^{\lambda}_{\text{NE}}$.
□ □

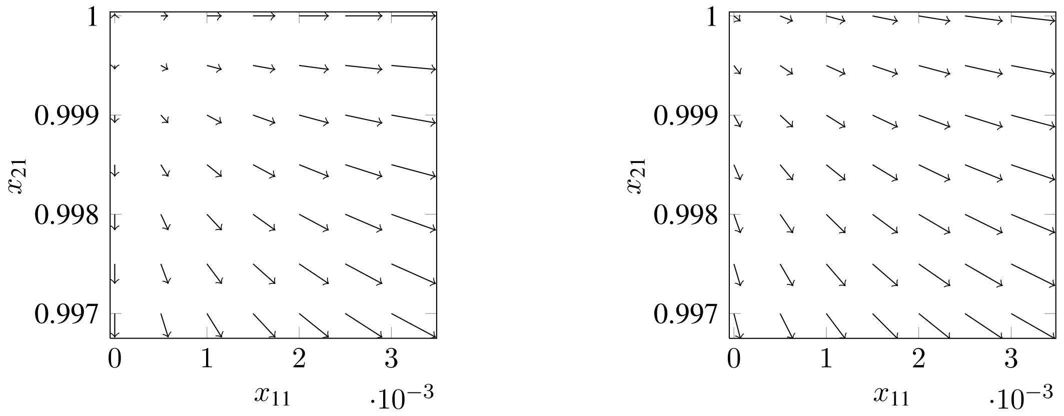
Note that due to the introduction of the state-dependent perturbation function in the decision rule of the players, vertices of $\mathbf{\Delta}$ cease to be stationary points of the ODE (12) when $\lambda > 0$. Due to this property, the introduction of the state-dependent perturbation function will address issues related to showing nonconvergence to boundary points which do not correspond to Nash equilibria [27, 4]. This will become more apparent in the forthcoming Section 6 when we discuss the probability of convergence to boundary points.

5.3 Demonstration

To demonstrate the sensitivity of the stationary points to the perturbation function, we plot the vector field of the ODE (12) in the vicinity of $\mathbf{\Delta}^*$, i.e., the vertices of the domain $\mathbf{\Delta}$. For demonstration purposes, we assume that there are two agents whose utility function is defined by Table 1, i.e., there are two pure Nash equilibria corresponding to action profiles (A, A) and (B, B).

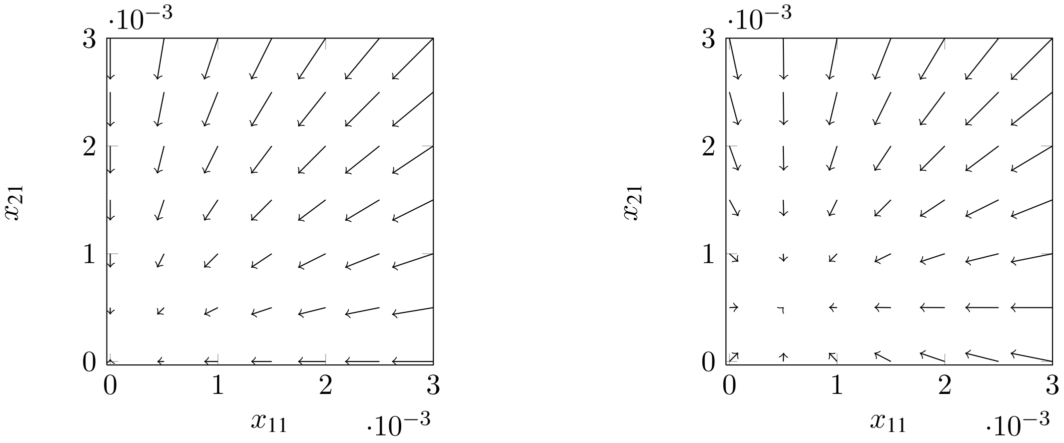
Fig. 2 plots the vector field of the ODE (12) in the vicinity of a non-Nash pure strategy profile, specifically in the vicinity of (B, A) which corresponds to strategies $1 - x^*_{11} = x^*_{21} = 1$. We observe that this is a stationary point of the ODE (12) for $\lambda = 0$, while it is no longer a stationary point when $\lambda = 0.01$. This conclusion agrees with the second statement of Lemma 5.2 which states that for sufficiently small neighborhood $\mathcal{B}_{\delta}(x^*)$ of a non-Nash action profile x^*, and for sufficiently small $\lambda > 0$, $\mathcal{B}_{\delta}(x^*)$ does not contain any stationary point of the ODE (12).

On the other hand, Fig. 3 plots the vector field of the ODE (12) in the vicinity of (B, B) which corresponds to strategies $x^*_{11} = x^*_{21} = 0$. As expected, when $\lambda = 0$, this strategy allocation corresponds to a stationary point of the ODE as shown in Fig. 3(a). When, instead, $\lambda = 0.01$ in Fig. 3(b), observe the slight displacement of the original stationary point towards the interior of the probability simplex as predicted by the first statement of Lemma 5.2.



(a) (b)

Figure 2: Sensitivity of a non-Nash stationary point to λ: (a) $\lambda = 0$, (b) $\lambda = 0.01$ and $\beta = 0.9$.



(a) (b)

Figure 3: Sensitivity of a strict Nash stationary point to λ: (a) $\lambda = 0$, (b) $\lambda = 0.01$ and $\beta = 0.9$.

6 Convergence to Boundary Points

Recall that, for the unperturbed dynamics, not all stationary points in $\mathbf{\Delta}^*$ are necessarily Nash equilibria. Convergence to non-desirable stationary points, such as the ones which are not Nash equilibria, cannot be excluded when agents employ the unperturbed reinforcement scheme $\widetilde{\mathcal{L}}_{\mathrm{R-I}}$.

Proposition 6.1 (Convergence to boundary points) *Under the reinforcement scheme $\widetilde{\mathcal{L}}_{\mathrm{R-I}}$, the probability that the same action profile will be played for all future times is uniformly bounded away from zero over all initial conditions if $R_i(\alpha) > 1$ for each $\alpha \in \mathcal{A}$, $i \in \mathcal{I}$.*

Proof. See Appendix B. □

Proposition 6.1 reveals the main issue of applying reinforcement learning schemes, which is convergence with positive probability to boundary points which are not Nash equilibrium profiles.

Fig. 4 shows a typical response of $\widetilde{\mathcal{L}}_{\mathrm{R-I}}$ in the Typewriter Game of Table 1. We observe that it is possible for the process to converge to a pure strategy profile which is not a Nash equilibrium when $R_i(\alpha) > 1$ for all $\alpha \in \mathcal{A}$ and $i \in \mathcal{I}$.

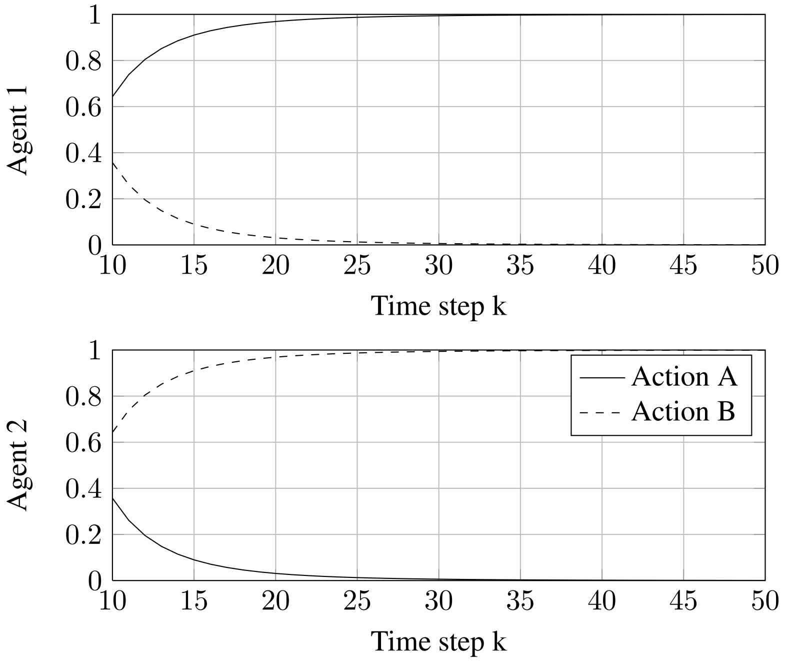


Figure 4: Typical response of $\widetilde{\mathcal{L}}_{\mathrm{R-I}}$ on the Typewriter Game of Table 1 when $\nu = 0.78$.

This issue has been addressed before in references [27, 4]. In particular, in [27] the reinforcement model of [1] was considered. The only difference of the learning model of [1] with $\widetilde{\mathcal{L}}_{\mathrm{R-I}}$ is the step-size sequence, which in case of [1] is defined as $\epsilon(k) = 1/(ck^{\nu} + R_i(\alpha))$ for some positive parameter c and for $0 < \nu < 1$. Reference [27] showed that convergence to pure strategy profiles which are not Nash equilibria can be excluded as long as $c > R_i(\alpha)$ for all $i \in \mathcal{I}$ and $\nu = 1$. This statement agrees with Proposition 6.1, since in $\widetilde{\mathcal{L}}_{\mathrm{R-I}}$ we have $c = 1$.

An alternative approach for guaranteeing nonconvergence to stationary points which are not Nash equilibria is the urn process of [3]. This model can be rewritten in the recursive form of $\widetilde{\mathcal{L}}_{\mathrm{R-I}}$ for which the step-size sequence will be $\epsilon_i(k) = 1/(V_i(k) + R_i(\alpha))$, where $V_i(k)$ is the accumulated benefits of agent i up to time k. This model has been analyzed in [4], where it was shown that the recursion converges with probability zero to any stationary point of the replicator dynamics which is not a Nash equilibrium. However, as [4] points out and we also showed in Proposition 6.1, similar statements cannot be derived for more general reinforcement learning schemes.

The *perturbed* reinforcement scheme $\widetilde{\mathcal{L}}^{\lambda}_{\mathrm{R-I}}$ introduced in Section 3 will provide an alternative approach for dealing with nonconvergence to pure-strategy profiles which are not Nash equilibria and will allow for establishing a connection to standard replicator dynamics (13).

7 Convergence of Perturbed Dynamics ($\widetilde{\mathcal{L}}^{\lambda}_{\mathrm{R-I}}$)

The convergence analysis of the perturbed dynamics $\widetilde{\mathcal{L}}^{\lambda}_{\mathrm{R-I}}$ will be subject to the following assumption:

Assumption 7.1 *For the unperturbed dynamics, $\widetilde{\mathcal{L}}_{\mathrm{R-I}}$, there exists a twice continuously differentiable and nonnegative function $V : \mathbf{\Delta} \to \mathbb{R}_+$ such that*

(a) $\nabla_x V(x)^{\mathrm{T}} \overline{g}(x) \leq 0$ for all $x \in \mathbf{\Delta}$;

(b) $\nabla_x V(x)^{\mathrm{T}} \overline{g}(x) = 0$ if and only if $\overline{g}(x) = 0$.

For some $\delta > 0$, consider the δ-neighborhood of the set of stationary points $\mathcal{S}^{\lambda}$, $\mathcal{B}_{\delta}(\mathcal{S}^{\lambda})$. Define also the closed set: $\mathcal{D}_{\delta}(\mathcal{S}^{\lambda}) \triangleq \mathbf{\Delta} \backslash \mathcal{B}_{\delta}(\mathcal{S}^{\lambda})$.

Lemma 7.1 *Under Assumption 7.1, for $\beta \in (0,1)$ sufficiently close to one and $\lambda > 0$ sufficiently small, there exists $\delta = \delta(\beta, \lambda) > 0$ such that*

$$\sup_{x \in \mathcal{D}_{\delta}(\mathcal{S}^{\lambda})} \nabla_x V(x)^{\mathrm{T}} \overline{g}^{\lambda}(x) < 0.$$

Proof. Pick $\delta^* = \delta^*(\beta)$ according to Proposition 5.2, such that, for all $x \in \mathbf{\Delta} \backslash \mathcal{B}_{\delta^*}(\mathbf{\Delta}^*)$, $\zeta_i(x_i, \lambda) = 0$ for at least one agent i. Then, according to Proposition 5.2, when we take β sufficiently close to one (which implies that δ^* approaches zero) and λ sufficiently small, then (a) $\widetilde{\mathcal{S}}^{\lambda}_{\mathrm{NE}} \subset \mathcal{B}_{\delta^*}(\mathbf{\Delta}^*)$, and (b) $\mathcal{S}^{\lambda} = \mathcal{S}_{\mathbf{\Delta}^o} \cup \widetilde{\mathcal{S}}^{\lambda}_{\mathrm{NE}}$. Due to Assumption 7.1, there exists $\delta = \delta(\beta, \lambda) > \delta^*$ such that $\mathcal{B}_{\delta^*}(\mathbf{\Delta}^*) \subset \mathcal{B}_{\delta}(\mathcal{S}^{\lambda})$ and

$$\sup_{x \in \mathcal{D}_{\delta}(\mathcal{S}^{\lambda})} \nabla_x V(x)^{\mathrm{T}} \overline{g}^{\lambda}(x) < 0.$$

Thus, the conclusion follows. □ □

Lemma 7.2 (LAS - $\widetilde{\mathcal{L}}^{\lambda}_{\mathrm{R-I}}$) *For any $\lambda > 0$ sufficiently small, any stationary point $\tilde{x} \in \widetilde{\mathcal{S}}^{\lambda}_{\mathrm{NE}}$, which is a perturbation of a strict Nash equilibrium according to (15), is a locally asymptotically stable point of the ODE (12).*

Proof. The proof follows similar reasoning with the proof of Proposition 3.6 in [14]. □ □

Theorem 7.1 (Convergence to Nash equilibria) *Under Assumption 7.1, if agents employ the perturbed reinforcement scheme $\widetilde{\mathcal{L}}^{\lambda}_{\mathrm{R-I}}$ for some $\beta \in (0,1)$ sufficiently close to one and $\lambda > 0$ sufficiently small, then there exists $\delta = \delta(\beta, \lambda)$ such that,*

$$\mathbb{P}\Big[\liminf_{k \to \infty} \mathrm{dist}(x(k), \mathcal{B}_{\delta}(\mathcal{S}^{\lambda})) = 0\Big] = 1.$$

Also, for almost all ω, the process $\{\psi_k(\omega) = x(k)\}$ converges to some invariant set in $\mathcal{B}_{\delta}(\mathcal{S}^{\lambda})$.

Proof. Consider the nonnegative function $V(x)$ of Assumption 7.1. We can approximate the expected incremental gain of $V(x)$ by applying a Taylor series expansion as follows:

$$\Delta V(k,x) = \nabla_x V(x)^{\mathrm{T}} \mathbb{E}[x(k+1) - x(k) | x(k) = x] + O(\epsilon(k)^2),$$

where $O(\epsilon(k)^2)$ denotes terms of order $\epsilon(k)^2$. Note that such an expansion is possible due to the fact that the second-order derivatives of $V(\cdot)$ are continuous in $\mathbf{\Delta}$. Equivalently,

$$\Delta V(k,x) = \epsilon(k) \nabla_x V(x)^{\mathrm{T}} \overline{g}^{\lambda}(x) + O(\epsilon(k)^2). \tag{16}$$

Due to Lemma 7.1, there exists $\delta = \delta(\beta, \lambda) > 0$ such that

$$-\bar{\rho} \triangleq \sup_{x \in \mathcal{D}_\delta(\mathcal{S}^\lambda)} \nabla_x V(x)^{\mathrm{T}} \overline{g}^{\lambda}(x) < 0.$$

Thus,

$$\Delta V(k,x) \leq -\epsilon(k)\bar{\rho} + O(\epsilon(k)^2),$$

uniformly in $x \in \mathcal{D}_\delta(\mathcal{S}^\lambda)$. The right-hand side of the above inequality is strictly negative and can be formulated in the form of condition (9). Therefore, the conditions of Corollary 4.1 are satisfied and

$$\mathbb{P}\Big[\liminf_{k \to \infty} \operatorname{dist}(x(k), \mathcal{B}_\delta(\mathcal{S}^\lambda)) = 0 \Big] = 1.$$

From Proposition 4.2, we also have that the process $\{\psi_k(\omega) = x(k)\}$ will converge to some invariant set of the ODE in $\mathcal{B}_\delta(\mathcal{S}^\lambda)$ almost surely. □ □

Theorem 7.1 is an immediate implication of Assumptions 5.1–7.1 and Lemma 7.1 and shows that the perturbed reinforcement scheme $\widetilde{\mathcal{L}}^{\lambda}_{\mathrm{R-I}}$ converges almost surely to some invariant set within an arbitrarily small neighborhood of the stationary points $\mathcal{S}^\lambda$ of the perturbed dynamics.

Recall also that, according to Proposition 5.2, the set $\mathcal{S}^\lambda$ includes a) the stationary points in $\mathbf{\Delta}^o$ which are perturbations of $\mathcal{S}_{\mathbf{\Delta}^*} \cap \mathcal{S}_{\mathrm{NE}}$ (i.e., stationary points which correspond to *strict* and *non-strict* pure Nash equilibria), and b) interior stationary points of the unperturbed dynamics, $\mathcal{S}_{\mathbf{\Delta}^o}$ (e.g., mixed Nash equilibria). Although Theorem 7.1 does not explicitly characterize the invariant sets within $\mathcal{B}_\delta(\mathcal{S}^\lambda)$ to which convergence is attained, $\delta = \delta(\beta, \lambda)$ can become arbitrarily small by appropriately selecting β and λ.

If we further exclude convergence to $\mathcal{S}_{\mathbf{\Delta}^o}$ (which is possible for some classes of games as we shall see in the forthcoming Section 8), then Theorem 7.1 will imply convergence to an arbitrarily small neighborhood of $\widetilde{\mathcal{S}}^{\lambda}_{\mathrm{NE}}$. In this case, given that the stationary points in $\widetilde{\mathcal{S}}^{\lambda}_{\mathrm{NE}}$ can become arbitrarily close to the corresponding vertices of the simplex (due to Proposition 5.2(a)), Theorem 7.1 implicitly implies convergence to an arbitrarily small neighborhood of the corresponding vertices of the simplex (i.e., the ones corresponding to λ-perturbations of Nash equilibria of the game). We will discuss this observation in greater detail in the forthcoming Section 8.

8 Specialization to Potential Games

8.1 Potential games

In this section, we will specialize the convergence analysis of Section 7 to a class of games which belongs to the general family of *ordinal potential games* (cf., [15]). In particular, we will consider games which satisfy the following property:

Property 8.1 *There exists a C^2 function $f : \mathbf{\Delta} \to \mathbb{R}$ such that*

$$\nabla_{\sigma_i} f(\sigma) = U_i(\sigma)$$

for all $\sigma \in \mathbf{\Delta}$ and $i \in \mathcal{I}$.

This property has been used to define potential games in population games [28], where players from an infinite-size population are paired to play the game, and σ corresponds to the average strategy in the population. The model here is equivalent, since instead of an infinite-size population of players and finite strategies, we consider a finite number of players with a continuum of strategies. A straightforward calculation can show that the function f serves as a *potential function* under the definition of [15], since for every $i \in \mathcal{I}$ and $\sigma_i, \sigma_i' \in \Delta(|\mathcal{A}_i|)$, we have

$$\begin{aligned} f(\sigma_i', \sigma_{-i}) - f(\sigma_i, \sigma_{-i}) &= \nabla_{\sigma_i} f(\sigma)^{\mathrm{T}} (\sigma_i' - \sigma_i) \\ &= U_i(\sigma)^{\mathrm{T}} (\sigma_i' - \sigma_i) \\ &= u_i(\sigma_i', \sigma_{-i}) - u_i(\sigma_i, \sigma_{-i}) \end{aligned}$$

where the first equality results from the Taylor series expansion of f about $\sigma = (\sigma_i, \sigma_{-i}) \in \mathbf{\Delta}$ and the fact that $\nabla^2_{\sigma_i} f(\sigma) = 0$.

Example 1: (*Common-payoff games*) One class of games which satisfies Property 8.1 is *common-payoff* or *identical-interest games*, where the payoff function is the same for all players. In other words, there exists a function $d : \mathcal{A} \to \mathbb{R}_+$ such that the expected payoff of player $i \in \mathcal{I}$ at strategy profile σ is:

$$u_i(\sigma) = \sum_{\alpha \in \mathcal{A}} d(\alpha) \prod_{k \in \mathcal{I}} \sigma_{k\alpha_k}.$$

Define $f(\sigma) \triangleq u_i(\sigma)$ for some $i \in \mathcal{I}$. Then, it is straightforward to check that

$$\frac{\partial f(\sigma)}{\partial \sigma_{ij}} = \sum_{\{\alpha : \alpha_i = j\}} d(\alpha) \prod_{k \in -i} \sigma_{k\alpha_k} = U_{ij}(\sigma),$$

i.e., f satisfies Property 8.1. An example of a common-payoff game is the Typewriter Game of Table 1.

Example 2: (*Congestion games*) A typical congestion game consists of a set $\mathcal{I}$ of n players and a set $\mathcal{P}$ of m paths. For each player i, let the set of pure strategies $\mathcal{A}_i$ be the set of m paths. The cost to each player i of selecting the path p depends on the number of players that are using the same path. The expected number of players using path p is

$$\chi_p(\sigma) \triangleq \sum_{i \in \mathcal{I}} \sigma_{ip}.$$

Define $c_p = c_p(\chi_p)$ to be the cost of using path p when χ_p players are using path p and let $c_p(\chi_p)$ be linear on χ_p. The expected utility of player i is defined as:

$$u_i(\sigma) \triangleq - \sum_{p \in \mathcal{P}} c_p(\chi_p(\sigma)).$$

Note that the function

$$f(\sigma) \triangleq - \sum_{p \in \mathcal{P}} \int_0^{\chi_p(\sigma)} c_p(z) dz$$

satisfies Property 8.1.

8.2 Convergence to Nash equilibria

The following proposition establishes convergence to Nash equilibria for this class of potential games.

Proposition 8.1 (Convergence to Nash equilibria) *For the class of games satisfying Property 8.1, the perturbed reinforcement scheme $\widetilde{\mathcal{L}}_{\mathrm{R-I}}^{\lambda}$ satisfies the conclusions of Theorem 7.1.*

Proof. It suffices to show that the conditions of Assumption 7.1 are satisfied for the unperturbed dynamics. In particular, define the nonnegative function

$$V(x) \triangleq f_{\max} - f(x) \geq 0, \quad x \in \mathbf{\Delta}, \tag{17}$$

where $f_{\max} \triangleq \sup_{x \in \mathbf{\Delta}} f(x)$. Note that $\nabla_{x_i} V(x) = -U_i(x)$, and

$$\begin{aligned} U_i(x)^{\mathrm{T}} \overline{g}_i(x) &= U_i(x)^{\mathrm{T}} X_i(x_i) U_i(x) \\ &= \sum_{s=1}^{|\mathcal{A}_i|} \sum_{j=1, j>s}^{|\mathcal{A}_i|} x_{is} x_{ij} (U_{is}(x) - U_{ij}(x))^2 \\ &\geq 0. \end{aligned}$$

Thus,

$$\nabla_x V(x)^{\mathrm{T}} \overline{g}(x) = -U(x)^{\mathrm{T}} \overline{g}(x) = - \sum_{i \in \mathcal{I}} U_i(x)^{\mathrm{T}} X_i(x_i) U_i(x) \leq 0$$

for all $x \in \mathbf{\Delta}$.

We also observe that $\nabla_x V(x)^{\mathrm{T}} \overline{g}(x) = 0$ if and only if $U_{is}(x) = U_{ij}(x)$ for any $i \in \mathcal{I}$ and any $s, j \in \mathcal{A}_i$, $s \neq j$ such that $x_{is}, x_{ij} > 0$. By Proposition 5.1, these points correspond to the stationary points of $\overline{g}(x)$. Therefore, the conditions of Assumption 7.1 are also satisfied. Thus, the conclusions of Theorem 7.1 hold for the class of games satisfying Property 8.1. □ □

8.3 Convergence to pure Nash equilibria

In several games, convergence to mixed Nash equilibria of the unperturbed dynamics $\mathcal{S}_{\mathbf{\Delta}^o}$ can be excluded. In this case, convergence of the perturbed dynamics to stationary points in $\widetilde{\mathcal{S}}_{\mathrm{NE}}^{\lambda}$ which are perturbations of pure Nash equilibria can be established.

Let x_{-i} denote the distribution over action profiles of the group of agents $-i$. Let D_i be the matrix of payoffs of agents i and D_{-i} be the matrix of payoffs of $-i$. The vector of expected payoffs of agent i and $-i$ can be expressed as $U_i(x) = D_i x_{-i}$ and $U_{-i}(x) = D_{-i} x_i$, respectively.

To analyze the behavior around stationary points in $\mathbf{\Delta}^o$, we consider the nonnegative function $V(x) \triangleq f_{\max} - f(x) \geq 0$, $x \in \mathbf{\Delta}$, where $f_{\max} \triangleq \sup_{x \in \mathbf{\Delta}} f(x)$. It is straightforward to verify that the Jacobian matrix of $f(x)$ is:

$$\nabla_x^2 f(x) = \begin{pmatrix} O & D_i \\ D_{-i} & O \end{pmatrix}.$$

Higher-order derivatives of $f(x)$ will be zero, therefore from the extension of Taylor's Theorem (cf., Theorem 5.15 in [29]) to multivariable functions, we have:

$$\begin{aligned} \Delta V(k,x) = -\nabla_x f(x)^{\mathrm{T}} \mathbb{E}[\delta x(k) | x(k) = x] - \\ \mathbb{E}[\delta x_{-i}(k)^{\mathrm{T}} D_{-i} \delta x_i(k) | x(k) = x] - \\ \mathbb{E}[\delta x_i(k)^{\mathrm{T}} D_i \delta x_{-i}(k) | x(k) = x], \end{aligned} \tag{18}$$

where $\delta x(k) \triangleq x(k+1) - x(k)$.

A direct consequence of the above formulation and Proposition 4.1 is the following:

Proposition 8.2 (Nonconvergence to $\mathcal{S}_{\mathbf{\Delta}^o}$) *If agents employ the unperturbed reinforcement scheme $\widetilde{\mathcal{L}}_{\mathrm{R-I}}$ and $x^* \in \mathcal{S}_{\mathbf{\Delta}^o}$ satisfies*

1. $\mathbb{E}[\delta x_{-i}(k)^{\mathrm{T}} D_{-i} \delta x_i(k) | x(k) = x] > 0$,
2. $\mathbb{E}[\delta x_i(k)^{\mathrm{T}} D_i \delta x_{-i}(k) | x(k) = x] > 0$

uniformly in $x \in \mathcal{B}_\delta(x^*)$, for some $\delta > 0$ sufficiently small, then

$$\mathbb{P}\Big[\lim_{k \to \infty} x(k) = x^*\Big] = 0.$$

Proof. We consider the nonnegative function $V(x)$ defined above. Note that the expected incremental gain of $V(x)$ (18), under the unperturbed dynamics, can take the following form:

$$V(k,x) = -\epsilon(k)\phi(k,x)$$

where $\inf_{x \in \mathcal{B}_\delta(x^*)} \phi(k,x) > 0$ for some $\delta > 0$ sufficiently small and for all k. This is due to the fact that for any $x \in \mathcal{B}_\delta(x^*)$,

$$-\nabla_x f(x)^{\mathrm{T}} \mathbb{E}[\delta x(k) | x(k) = x] \leq 0$$

(due to Proposition 8.1), and the second-order terms of the incremental gain are strictly negative by assumption. Then, from Proposition 4.1, we conclude that the unperturbed process will exit $\mathcal{B}_\delta(x^*)$ in finite time with probability one. Therefore, the conclusion follows. □ □

For several games testing the conditions of Proposition 8.2 may be hard. However, for two-player two-action games, it is straightforward to show that:

$$\begin{aligned} &\mathbb{E}[\delta x_i{}^{\mathrm{T}} D_i \delta x_{-i} | x_i(k) = x_i, x_{-i}(k) = x_{-i}] = \\ &\quad \epsilon(k)^2 x_{i1} x_{i2} x_{(-i)1} x_{(-i)2} (d^i_{11} - d^i_{12} - d^i_{21} + d^i_{22}) \cdot \\ &\quad ((d^i_{11})^2 - (d^i_{12})^2 - (d^i_{21})^2 + (d^i_{22})^2), \end{aligned} \tag{19}$$

where $d^i_{s\ell}$ denotes the (s, ℓ) entry of D_i, $i = 1, 2$. Consider, for example, the Typewriter Game of Table 1. Since the game is symmetric, and $d^i_{11} > d^i_{12}$, $d^i_{22} > d^i_{21}$, $i = 1, 2$, the second-order terms of the incremental gain will be positive. The above computation can be extended in a similar manner to the case of larger number of actions.

Proposition 8.3 (Convergence to pure Nash equilibria) *In the framework of Proposition 8.1, let the conditions of Proposition 8.2 also hold. If the game admits pure Nash equilibria which are all strict, then, for some $\beta \in (0, 1)$ sufficiently close to one and $\lambda > 0$ sufficiently small, the perturbed process $\{\psi_k(\omega) = x(k)\}$ converges to the set $\widetilde{\mathcal{S}}^\lambda_{\mathrm{NE}}$ for almost all ω, i.e.,*

$$\mathbb{P}\Big[\lim_{k\to\infty} x(k) \in \widetilde{\mathcal{S}}^\lambda_{\mathrm{NE}}\Big] = 1.$$

Proof. Since the game exhibits pure Nash equilibria which are all strict, the set $\widetilde{\mathcal{S}}^\lambda_{\mathrm{NE}}$ in non-empty for any $\lambda > 0$ sufficiently small.

Let x^* denote an action profile which is a strict pure Nash equilibrium, i.e., for every $i \in \mathcal{I}$ there exists $j^* = j^*(i)$ such that $x_{ij^*} = 1$ and $U_{is}(x^*) - U_{ij^*}(x^*) < 0$ for any $s \neq j^*$. Let also $\tilde{x} \in \widetilde{\mathcal{S}}^\lambda_{\mathrm{NE}}$ be the perturbed stationary point according to (15). Pick also $\delta^* = \delta^*(\beta) > 0$ similarly to the proof of Lemma 7.1. Then, for any $x \in \mathcal{B}_{\delta^*}(\tilde{x})$, x_{is} is of order of δ^* and

$$\overline{g}^\lambda_{is}(x) \approx [U_{is}(x^*) - U_{ij^*}(x^*)]\, x_{is} \tag{20}$$

plus higher order terms of δ^* and λ, for all $s \neq j^*$. Since $U_{is}(x^*) - U_{ij^*}(x^*) < 0$ for all $s \neq j^*$, we conclude that the vector field points towards the interior of $\mathcal{B}_{\delta^*}(\tilde{x})$ when evaluated at the boundary of $\mathcal{B}_{\delta^*}(\tilde{x})$. Thus, $\mathcal{B}_{\delta^*}(\tilde{x})$ is an invariant set of the ODE (12). Therefore, due to Proposition 8.2 and Theorem 7.1, if we take $\beta \in (0, 1)$ sufficiently close to one and $\lambda > 0$ sufficiently small, then there exists $\delta = \delta(\beta, \lambda) > \delta^*$ such that the process $\{x(k)\}$ converges almost surely to some invariant set in $\mathcal{B}_\delta(\widetilde{\mathcal{S}}^\lambda_{\mathrm{NE}})$.

Furthermore, due to Lemma 7.2, we know that the points in $\widetilde{\mathcal{S}}^\lambda_{\mathrm{NE}}$ are locally asymptotically stable, and therefore by (20), the set $\mathcal{B}_\delta(\widetilde{\mathcal{S}}^\lambda_{\mathrm{NE}})$ belongs to its region of attraction. Since the perturbed process visits $\mathcal{B}_\delta(\widetilde{\mathcal{S}}^\lambda_{\mathrm{NE}})$ infinitely often, by Proposition 4.2, we conclude that the process converges to $\widetilde{\mathcal{S}}^\lambda_{\mathrm{NE}}$ with probability one. □ □

Proposition 8.3 specializes the conclusions of Theorem 7.1 to the case where i) Property 8.1 is satisfied, ii) the pure Nash equilibria of the game are all strict, and iii) the hypotheses of Proposition 8.2 also hold (i.e., convergence to mixed Nash equilibria can be excluded). Proposition 8.3 shows that asymptotic convergence to the set of stationary points $\widetilde{\mathcal{S}}_{\mathrm{NE}}^{\lambda}$ (i.e., λ-perturbations of pure Nash equilibria) can be achieved almost surely.

In case the game exhibits pure Nash equilibria which are *not* strict, the conclusions of Proposition 8.3 might not hold. However, Theorem 7.1 still applies. In particular, under the hypotheses of Proposition 8.2, Theorem 7.1 implies that the perturbed process will converge almost surely to an invariant set within an arbitrarily small neighborhood of $\widetilde{\mathcal{S}}_{\mathrm{NE}}^{\lambda}$ (i.e., λ-perturbations of pure Nash equilibria). Furthermore, due to Proposition 5.2, the stationary points $\widetilde{\mathcal{S}}_{\mathrm{NE}}^{\lambda}$ can become arbitrarily close to the corresponding vertices of the simplex by appropriately selecting parameters β and λ. We conclude that, even if the game exhibits pure Nash equilibria which are not strict, Theorem 7.1 implies that the perturbed process converges almost surely to an invariant set within an arbitrarily small neighborhood of the vertices corresponding to $\widetilde{\mathcal{S}}_{\mathrm{NE}}^{\lambda}$.

8.4 Extension to Two-Player Rescaled Partnership Games

In two-player games, the convergence results of Propositions 8.1–8.3 can be extended to two-player rescaled partnership games, introduced by [10] and defined as follows:

Definition 8.1 (Two-Player Rescaled Partnership Games) *A two-player game with payoff matrices D_i, $i \in \{1,2\}$, is a rescaled partnership game if there exist positive numbers a_i, $i \in \{1,2\}$, and matrices*

$$C_i = \begin{pmatrix} c_{i1} & c_{i2} \\ c_{i1} & c_{i2} \end{pmatrix} \in \mathbb{R}^{2\times 2}, \quad i \in \mathcal{I},$$

such that the two-player game with payoff matrices

$$D_i' \triangleq a_i D_i + C_i, \quad i \in \{1,2\},$$

define a partnership game, i.e., $D_i' = \left(D_{-i}'\right)^{\mathrm{T}}$.

Note that two-player partnership games are also potential games with potential function $f : \boldsymbol{\Delta} \to \mathbb{R}$ such that

$$f(\sigma) \triangleq \sigma_i^{\mathrm{T}} D_i' \sigma_{-i}, \tag{21}$$

for some $i \in \{1,2\}$.

As already pointed out by [4], two-player rescaled partnership games exhibit a nice property in connection with standard replicator dynamics, summarized in the following claim.

Claim 8.1 *For any two-player rescaled partnership game and for any $x \in \boldsymbol{\Delta}$, the following holds:*

$$X_i(x_i) D_i' x_{-i} = a_i X_i(x_i) D_i x_{-i}, \quad i \in \{1,2\}. \tag{22}$$

Due to this property, convergence to Nash equilibria in rescaled partnership games can be established under the perturbed reinforcement scheme $\widetilde{\mathcal{L}}_{\mathrm{R-I}}^{\lambda}$.

Proposition 8.4 (Convergence in Rescaled Partnership Games) *In the class of two-player rescaled partnership games, the $\widetilde{\mathcal{L}}_{\mathrm{R-I}}^{\lambda}$ reinforcement scheme satisfies the conclusions of Theorem 7.1. Furthermore, if the conditions of Proposition 8.2 hold and the game admits pure Nash equilibria which are all strict, then, for some $\beta \in (0,1)$ sufficiently close to one and $\lambda > 0$ sufficiently small, the process $\{\psi_k(\omega) = x(k)\}$ converges to the set $\widetilde{\mathcal{S}}_{\mathrm{NE}}^{\lambda}$ for almost all ω, i.e.,*

$$\mathbb{P}\Big[\lim_{k\to\infty} x(k) \in \widetilde{\mathcal{S}}_{\mathrm{NE}}^{\lambda}\Big] = 1.$$

Proof. It suffices to show that the conditions of Assumption 7.1 are satisfied. In particular, define the nonnegative function

$$V(x) \triangleq f_{\max} - f(x) \geq 0, \quad x \in \mathbf{\Delta},$$

where f is defined according to (21) and $f_{\max} \triangleq \sup_{x\in\mathbf{\Delta}} f(x)$. Note that $\nabla_{x_i} V(x) = -D_i' x_{-i}$, and

$$\begin{aligned}
\nabla_x V(x)^{\mathrm{T}} \bar{g}(x) &= -\sum_{i\in\mathcal{I}} x_{-i}^{\mathrm{T}} (D_i')^{\mathrm{T}} \bar{g}_i(x) \\
&= -\sum_{i\in\mathcal{I}} x_{-i}^{\mathrm{T}} (D_i')^{\mathrm{T}} X_i(x_i) D_i x_{-i} \\
&= -\sum_{i\in\mathcal{I}} a_i x_{-i}^{\mathrm{T}} D_i^{\mathrm{T}} X_i(x_i) D_i x_{-i} \\
&\leq 0
\end{aligned}$$

for all $x \in \mathbf{\Delta}$, where the last equality is due to property (22) and the last inequality is due to the fact that $X_i(x_i)$ is a positive semidefinite (symmetric) matrix (as was shown in the proof of Proposition 8.1 and has pointed out in Exercise 9.6.3 of [10]).

Also, due to (22), we have that the stationary points of the mean-field dynamics in a rescaled partnership game with payoff matrices D_i, $i \in \mathcal{I}$, coincide with the stationary points of the mean-field dynamics of the partnership game $D_i' = a_i D_i + C_i$, $i \in \mathcal{I}$. We also observe that

$$\nabla_x V(x)^{\mathrm{T}} \bar{g}(x) = 0 \quad \Leftrightarrow \quad a_i x_{-i}^{\mathrm{T}} D_i^{\mathrm{T}} X_i(x_i) D_i x_{-i} = 0, \quad \forall i \in \mathcal{I}$$

which, according to the proof of Proposition 8.1, is satisfied if and only if $x \in \mathbf{\Delta}$ is a stationary point of $\bar{g}(x)$.

Thus, the conditions of Assumption 7.1 are satisfied for the two-player rescaled partnership games and therefore the conclusions of Theorem 7.1 hold. Furthermore, if the conditions of Proposition 8.2 apply and the game admits pure Nash equilibria which are all strict, then the conclusions of Proposition 8.3 also hold. □ □

An analogous result to Proposition 8.4 has been shown by [4] for rescaled partnership games under the reinforcement learning scheme of [3]. Proposition 8.4 can be thought of as an extension to a larger class of reinforcement learning schemes (beyond the urn process of [3]), due to the freedom in the selection of the step-size sequence (4). In fact, analogous results under a perturbed decision rule of the form (6) can be derived in a straightforward manner for other forms of reinforcement learning schemes, e.g., the reinforcement scheme of [1].

9 Conclusions

This paper presented a novel reinforcement learning scheme for distributed convergence to Nash equilibria. The main difference from prior schemes lies in the introduction of a perturbation function in the decision rule of each agent which depends only on its own strategy. The introduction of such perturbation function sidestepped issues regarding the behavior of the algorithm close to the vertices of the simplex. In particular, we derived conditions under which the perturbed reinforcement learning scheme converges to an arbitrarily small neighborhood of the set of Nash equilibria almost surely. This constitutes our main contribution, since prior convergence analysis on reinforcement learning has primarily focused on urn processes. We further specialized the results to a class of games which belongs to the family of potential games. We finally extended the convergence results to two-player rescaled partnership games, where we derived conditions under which convergence to perturbations of strict pure Nash equilibria can be achieved.

A Proof of Proposition 5.2

For any agent $i \in \mathcal{I}$ and any action $s \in \mathcal{A}_i$, the corresponding entry of the vector field is

$$\overline{g}_{is}(\tilde{x}) = U_{is}(\tilde{x})[(1-\zeta_i)\tilde{x}_{is} + \zeta_i/\left|\mathcal{A}_i\right|] - \sum_{q\in\mathcal{A}_i} U_{iq}(\tilde{x})[(1-\zeta_i)\tilde{x}_{iq} + \zeta_i/\left|\mathcal{A}_i\right|]\tilde{x}_{is}, \tag{23}$$

where $\zeta_i = \zeta_i(\tilde{x}_i, \lambda)$.

Consider any pure strategy profile x^*, and take $\tilde{x} = x^* + \nu$, for some $\nu = (\nu_1, \nu_2, ..., \nu_n) \in \times_{i\in\mathcal{I}}\mathbb{R}^{|\mathcal{A}_i|}$ such that $\nu_i \in \text{null}\{\mathbf{1}^{\mathrm{T}}\}$ and $\tilde{x}_i = x_i^* + \nu_i \in \Delta(|\mathcal{A}_i|)$ for all $i \in \mathcal{I}$. Substituting $\tilde{x}$ into (23), yields

$$\begin{aligned}\overline{g}_{is}(\nu,\lambda) = {} & U_{is}(\tilde{x})\left[(1-\zeta_i(x_{is}^* + \nu_{is}) + \zeta_i/\left|\mathcal{A}_i\right|\right] \\ & - \sum_{q\in\mathcal{A}_i} U_{iq}(\tilde{x})\left[(1-\zeta_i)(x_{iq}^* + \nu_{iq}) + \zeta_i/\left|\mathcal{A}_i\right|\right](x_{is}^* + \nu_{is}).\end{aligned}$$

where $\zeta_i = \zeta_i(x_i^* + \nu_i, \lambda)$. The perturbation function has the following properties:

$$\left.\frac{\partial\zeta_i(\nu_i,\lambda)}{\partial\nu_{ij}}\right|_{(0,0)} = 0, \quad \text{for all } j \in \mathcal{A}_i.$$

Furthermore, $\overline{g}_{is}(0,0) = 0$, since x^* is a stationary point of the unperturbed dynamics. Thus, the partial derivatives of $\overline{g}_{is}$ evaluated at $(0,0)$ are:

$$\left.\frac{\partial \overline{g}_{is}(\nu,\lambda)}{\partial \nu_{is}}\right|_{(0,0)} = U_{is}(x^*)(1-x^*_{is}) - \sum_{q\in\mathcal{A}_i} U_{iq}(x^*)x^*_{iq},$$

$$\left.\frac{\partial \overline{g}_{is}(\nu,\lambda)}{\partial \nu_{iq}}\right|_{(0,0)} = -U_{iq}(x^*)x^*_{is}, \quad \text{for all } q\in\mathcal{A}_i\backslash s.$$

Note also that for any $\ell\in\mathcal{I}\backslash i$ and $m\in\mathcal{A}_\ell$, we have

$$\left.\frac{\partial \overline{g}_{is}(\nu,\lambda)}{\partial \nu_{\ell m}}\right|_{(0,0)} = \left.\frac{\partial U_{is}(\tilde{x})}{\partial \nu_{\ell m}}\right|_{(0,0)} x^*_{is} - \sum_{q\in\mathcal{A}_i} \left.\frac{\partial U_{iq}(\tilde{x})}{\partial \nu_{\ell m}}\right|_{(0,0)} x^*_{iq}x^*_{is}.$$

Since x^* corresponds to a pure strategy state, for each $i\in\mathcal{I}$ there exists $j^*=j^*(i)$ such that $x^*_i=e_{j^*}$, i.e., $x_{ij^*}=1$ and $x^*_{is}=0$ for all $s\neq j^*$. For this pure strategy state and for any $s\in\mathcal{A}_i\backslash j^*$ we have

$$\left.\frac{\partial \overline{g}_{is}(\nu,\lambda)}{\partial \nu_{is}}\right|_{(0,0)} = U_{is}(x^*) - U_{ij}(x^*),$$

and

$$\left.\frac{\partial \overline{g}_{is}(\nu,\lambda)}{\partial \nu_{iq}}\right|_{(0,0)} = 0 \quad \forall q\in\mathcal{A}_i\backslash s, \quad \left.\frac{\partial \overline{g}_{is}(\nu,\lambda)}{\partial \nu_{\ell m}}\right|_{(0,0)} = 0 \quad \forall \ell\in\mathcal{I}\backslash i, m\in\mathcal{A}_\ell.$$

Given that $\nu_i\in\text{null}\{\mathbf{1}^{\mathrm{T}}\}$ and $\partial\overline{g}_{is}(\nu,\lambda)/\partial\nu_{ij^*}=0$ for all $s\neq j^*$, the behavior of $\overline{g}(\cdot,\cdot)$ with respect to ν about the point $(0,0)$ is described by the following Jacobian matrix:

$$\nabla_\nu \overline{g}(\nu,\lambda)|_{(0,0)} = \begin{pmatrix} \text{diag}\{U_{1s}(x^*)-U_{1j^*}(x^*)\}_{s\neq j^*} & & 0 \\ & \ddots & \\ 0 & & \text{diag}\{U_{ns}(x^*)-U_{nj^*}(x^*)\}_{s\neq j^*} \end{pmatrix}.$$

The above Jacobian matrix has full rank if for each $i\in\mathcal{I}$

$$U_{is}(x^*) - U_{ij^*}(x^*) \neq 0 \quad \text{for all } s\neq j^*.$$

In this case, by the implicit function theorem, there exists a neighborhood D of $\lambda=0$ and a unique differentiable function $\nu^*: D\to\mathbb{R}^{|\mathcal{A}|}$ such that $\nu^*(0)=0$ and $\overline{g}(\nu^*(\lambda),\lambda)=0$, for any $\lambda\in D$.

To characterize exactly the stationary points for small values of λ, we need to also compute the gradient of the mean-field with respect to the perturbation parameter λ. Note that:

$$\left.\frac{\partial g_{is}(\nu,\lambda)}{\partial\lambda}\right|_{(0,0)} = \frac{U_{is}(\tilde{x})}{|\mathcal{A}_i|}\left.\frac{\partial\zeta_i}{\partial\lambda}\right|_{(0,0)} = \frac{U_{is}(\tilde{x})}{|\mathcal{A}_i|},$$

since the partial derivative of ζ_i with respect to λ when evaluated at $(0,0)$ is 1.

Thus,

$$\nabla_{\lambda}\overline{g}(\nu,\lambda)|_{(0,0)} = \left(\begin{array}{c} \mathrm{col}\{U_{1s}(x^*)/|\mathcal{A}_1|\}_{s\neq j^*} \\ \vdots \\ \mathrm{col}\{U_{ns}(x^*)/|\mathcal{A}_n|\}_{s\neq j^*} \end{array}\right).$$

Again, by implicit function theorem, we have that

$$\nabla_{\lambda}\nu^*(\lambda)|_{\lambda=0} = -(\nabla_{\nu}\overline{g}(\nu,\lambda)|_{(0,0)})^{-1}\cdot\nabla_{\lambda}\overline{g}(\nu,\lambda)|_{(0,0)}$$

which implies that for any $i\in\mathcal{I}$ and for any $s\neq j^*$

$$\left.\frac{d\nu_{is}^*(\lambda)}{d\lambda}\right|_{\lambda=0} = -\frac{1}{(U_{is}(x^*)-U_{ij^*}(x^*))}.$$

Since $\nu_{is}^*(0)=0$ and $x_{is}^*=0$, in order for the solution $\tilde{x}=x^*+\nu^*(\lambda)$ to be in $\mathbf{\Delta}^o$, we also need the condition $d\nu_{is}^*(\lambda)/d\lambda|_{\lambda=0}>0$ to be satisfied for all $s\neq j^*$. Since $U_{is}(x^*)>0$ by Assumption 3.1, this condition is equivalent to

$$U_{is}(x^*)-U_{ij^*}(x^*)<0$$

for all $i\in\mathcal{I}$ and any $s\neq j^*$. This is also equivalent to x^* being a strict Nash equilibrium. Thus, the conclusion follows.

If x^* corresponds to an action profile which is not a Nash equilibrium, then there exist $i\in\mathcal{I}$ and $s\neq j^*$ such that $U_{is}(x^*)-U_{ij^*}(x^*)>0$. For any $\beta\in(0,1)$ which is sufficiently close to one, there exist $\delta_0=\delta_0(\beta)$ such that $\zeta_i(x_i,\lambda)\equiv 0$, $i\in\mathcal{I}$, for any $x\in\mathbf{\Delta}\backslash\mathcal{B}_{\delta}(x^*)$, $\lambda>0$ and $\delta\geq\delta_0$. For any $x\in\mathcal{B}_{\delta}(x^*)$, $\delta\geq\delta_0$, the vector field becomes

$$\overline{g}_{is}^{\lambda}(x)\approx[U_{is}(x)-U_{ij^*}(x)]x_{is}+U_{is}(x)\zeta_i(x_i,\lambda)/|\mathcal{A}_i| \tag{24}$$

plus higher order terms of λ and δ, for all $s\neq j^*$. Since the Nash condition is violated in the direction of s, $U_{is}(x)-U_{ij^*}(x)=c+O(\delta)$, for some $c>0$, where $O(\delta)$ denotes a quantity of order of δ. Furthermore, by Assumption 3.1 of strictly positive rewards, $U_{is}(x)>0$ for all $s\in\mathcal{A}_i$ and $x\in\mathcal{B}_{\delta}(x^*)$. Therefore, for any $\delta\geq\delta_0$ and for sufficiently small $\lambda>0$, the vector field $\overline{g}_{is}(x)>0$ for any $x\in\mathcal{B}_{\delta}(x^*)$, which implies that there is no stationary point of the vector field in $\mathcal{B}_{\delta}(x^*)$.

B Proof of Proposition 6.1

Let us assume that action profile $\alpha=(\alpha_1,\alpha_2,...\alpha_n)\in\mathcal{A}$ has been selected at time $k=0$. This implies that $x_{i\alpha_i}(0)>0$, since actions are selected according to the strategy distribution $\sigma_i(0)=x_i(0)$. The corresponding payoff profile will be $R(\alpha)=(R_1(\alpha),R_2(\alpha),...,R_n(\alpha))$, where according to Assumption 3.1, $R_i(\alpha)>0$ for all $i\in\mathcal{I}$. Let us define the following event:

$$A_{\tau}\triangleq\{\omega\in\Omega:\psi_k(\omega)=\alpha(k)=\alpha\text{ for all }k\leq\tau\}.$$

Thus, A_τ corresponds to the case where the same action profile has been performed for all times $k \leq \tau$. Note that the sequence of events $\{A_\tau\}$ is decreasing, since $A_\tau \supseteq A_{\tau+1}$ for all $\tau = 1, 2, ...$ Define also the event

$$A_\infty \triangleq \bigcap_{\tau=1}^{\infty} A_\tau \equiv \{\alpha(\tau) = \alpha, \forall \tau\}.$$

Therefore, from continuity from above, we have:

$$\mathbb{P}[A_\infty] = \lim_{\tau\to\infty} \mathbb{P}[A_\tau] = \lim_{\tau\to\infty} \prod_{k=1}^{\tau} \prod_{i\in\mathcal{I}} x_{i\alpha_i}(k) \triangleq \chi(\alpha).$$

The above upper bound $\chi(\alpha)$ is non-zero if and only if

$$\sum_{k=1}^{\infty} \log(x_{i\alpha_i}(k)) > -\infty \text{ for each } i \in \mathcal{I}. \tag{25}$$

Let us define the new variable

$$y_i(k) \triangleq 1 - x_{i\alpha_i}(k) = \sum_{j\in\mathcal{A}_i\backslash\alpha_i} x_{ij}(k),$$

which corresponds to the probability of agent i selecting any action other than α_i. Equivalently, condition (25) is equivalent to

$$-\sum_{k=0}^{\infty} \log(1 - y_i(k)) < \infty, \quad \text{for each } i \in \mathcal{I}. \tag{26}$$

We also have that

$$\lim_{k\to\infty} \frac{-\log(1 - y_i(k))}{y_i(k)} = \lim_{k\to\infty} \frac{1}{1 - y_i(k)} > \rho$$

for some finite $\rho > 0$, since $0 \leq y_i(k) \leq 1$. Thus, from the limit comparison test, we conclude that condition (26) holds, if and only if

$$\sum_{k=1}^{\infty} y_i(k) < \infty, \quad \text{for each } i \in \mathcal{I}.$$

Since $\epsilon(k) = 1/(k^\nu + 1)$, for $1/2 < \nu \leq 1$, we have:

$$\frac{y_i(k+1)}{y_i(k)} = 1 - \frac{R_i(\alpha)}{k^\nu + 1} \leq 1 - \frac{R_i(\alpha)}{k+1}.$$

By Raabe's criterion, the series $\sum_{k=0}^{\infty} y_i(k)$ is convergent if

$$\lim_{k\to\infty} k\left(\frac{y_i(k)}{y_i(k+1)} - 1\right) > 1.$$

Since

$$k\left(\frac{y_i(k)}{y_i(k+1)} - 1\right) \geq k\left(\frac{1}{1 - \frac{R_i(\alpha)}{k+1}} - 1\right) = k\frac{R_i(\alpha)}{k + 1 - R_i(\alpha)} = \frac{R_i(\alpha)}{1 + \frac{1-R_i(\alpha)}{k}}$$

we conclude that the series $\sum_{k=0}^{\infty} y_i(k)$ is convergent if $R_i(\alpha) > 1$ for each $i \in \mathcal{I}$. In other words, the action profile α will be performed for all future times with positive probability if $R_i(\alpha) > 1$ for all $i \in \mathcal{I}$. Furthermore, if $R_i(\alpha) > 1$ for all $i \in \mathcal{I}$ and for all $\alpha \in \mathcal{A}$, then the probability that the same action profile will be played for all future times is uniformly bounded away from zero over all initial conditions.

References

[1] W. B. Arthur, “On designing economic agents that behave like human agents,” *Journal of Evolutionary Economics*, vol. 3, pp. 1–22, 1993.

[2] T. Börgers and R. Sarin, “Learning through reinforcement and replicator dynamics,” *Journal of Economic Theory*, vol. 77, no. 1, pp. 1–14, 1997.

[3] I. Erev and A. Roth, “Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria,” *American Economic Review*, vol. 88, pp. 848–881, 1998.

[4] E. Hopkins and M. Posch, “Attainability of boundary points under reinforcement learning,” *Games and Economic Behavior*, vol. 53, pp. 110–125, 2005.

[5] A. Beggs, “On the convergence of reinforcement learning,” *Journal of Economic Theory*, vol. 122, pp. 1–36, 2005.

[6] B. Skyrms and R. Pemantle, “A dynamic model of social network formation,” *Proc. of the National Academy of Sciences of the USA*, vol. 97, pp. 9340–9346, 2000.

[7] P. Bonacich and T. Liggett, “Asymptotics of a matrix-valued markov chain arising in sociology,” *Stochastic Processes and Their Applications*, vol. 104, pp. 155–171, 2003.

[8] J. M. Smith, *Evolution and the Theory of Games*. Cambridge: Cambridge University Press, 1982.

[9] D. Leslie, “Reinforcement learning in games,” Ph.D. dissertation, School of Mathematics, University of Bristol, 2004.

[10] J. Hofbauer and K. Sigmund, *Evolution Games and Population Dynamics*. Cambridge, UK: Cambridge University Press, 1998.

[11] R. Bush and F. Mosteller, *Stochastic Models of Learning*. New York, NY: John Wiley and Sons, 1955.

[12] R. Pemantle, “Nonconvergence to unstable points in urn models and stochastic approximations,” *The Annals of Probability*, vol. 18, no. 2, pp. 698–712, 1990.

[13] J. Bergin and B. L. Lipman, “Evolution with state-dependent mutations,” *Econometrica*, vol. 64, no. 4, pp. 943–956, 1996.

[14] G. Chasparis and J. Shamma, “Distributed dynamic reinforcement of efficient outcomes in multiagent coordination and network formation,” *Dynamic Games and Applications*, vol. 2, no. 1, pp. 18–50, 2012.

[15] D. Monderer and L. Shapley, “Potential games,” *Games and Economic Behavior*, vol. 14, pp. 124–143, 1996.

[16] R. Rosenthal, “A class of games possessing pure-strategy Nash equilibria,” *International Journal of Game Theory*, vol. 2, no. 1, pp. 65–67, 1973.

[17] K. Savla and E. Frazzoli, “Game-theoretic learning algorithm for a spatial coverage problem,” in *47th Annual Allerton Conference on Communication, Control and Computing*, Allerton, 2010.

[18] E. Altman, Y. Hayel, and H. Kameda, “Evolutionary dynamics and potential games in non-cooperative routing,” in *WiOpt 2007*, Limassol, 2007.

[19] K. Narendra and M. Thathachar, *Learning Automata: An introduction.* Prentice-Hall, 1989.

[20] M. F. Norman, “On linear models with two absorbing states,” *Journal of Mathematical Psychology*, vol. 5, pp. 225–241, 1968.

[21] I. J. Shapiro and K. S. Narendra, “Use of stochastic automata for parameter self-organization with multi-modal performance criteria,” *IEEE Transactions on Systems Science and Cybernetics*, vol. 5, pp. 352–360, 1969.

[22] I. K. Cho and A. Matsui, “Learning aspiration in repeated games,” *Journal of Economic Theory*, vol. 124, pp. 171–201, 2005.

[23] M. B. Nevelson and R. Z. Hasminskii, *Stochastic Approximation and Recursive Estimation.* Providence, RI: American Mathematical Society, 1976.

[24] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications.* Springer-Verlag New York, Inc., 1997.

[25] W. H. Sandholm, *Population Games and Evolutionary Dynamics.* Cambridge, MA: The MIT Press, 2010.

[26] J. Weibull, *Evolutionary Game Theory.* Cambridge, MA: MIT Press, 1997.

[27] M. Posch, “Cycling in a stochastic learning algorithm for normal form games,” *Evolutionary Economics*, vol. 7, pp. 193–207, 1997.

[28] W. Sandholm, “Potential games with continuous player sets,” *Journal of Economic Theory*, vol. 97, pp. 81–108, 2001.

[29] W. Rudin, *Principles of Mathematical Analysis.* McGraw-Hill Book Company, 1964.